ASYMPTOTICALLY EFFICIENT STOCHASTIC APPROXIMATION; THE RM CASE

BY VÁCLAV FABIAN

Michigan State University

Anbar (1971) and, independently, Abdelhamid (1971) have shown that if the density \( g \) of the errors of estimates of function values is known, a transformation of observations leads to stochastic approximation methods which under mild conditions produce asymptotically efficient estimators (the first author considers the RM case, the second the RM and KW cases). This paper treats the RM case and shows that the same asymptotic result can be achieved without the knowledge of the density \( g \).

1. Introduction. We shall be concerned with the so called Robbins–Monro situation in stochastic approximation. In this situation the goal is to estimate a number \( \theta \) by observing unbiased estimates of function values of a function \( f \), which is negative on \( (-\infty, \theta) \) and positive on \( (\theta, +\infty) \). The original RM procedure (Robbins and Monro (1951)) was of the form \( X_{n+1} = X_n - a_n Y_n \), where \( Y_n \) have conditional, given \( X_1, X_2, \ldots, X_n \), expectations \( f(X_n) \) and bounded variances. Considerations of the speed showed that optimal constants \( a_n \) are of the form \( a_n = an^{-1} \) with \( a = (f'(\theta))^{-1} \) and that the procedure can be modified in such a way that, with \( f'(\theta) \) unknown, it has the same asymptotic properties as the original procedure with the optimal constant \( a = (f'(\theta))^{-1} \). This result was obtained by Venter (1967) and generalized by Fabian (1968). Under mild conditions and if the conditional variance of \( Y_n \), given the past \( X_1, \ldots, X_n \), converges to \( \sigma^2 \) if \( n \to \infty \) and \( X_n \to \theta \), the result is that \( X_n \to \theta \) a.e. and \( n^{-1}(X_n - \theta) \) are asymptotically normal \( (0, \sigma^2/(f'(\theta))^2) \). The indicated variance is easily seen minimal in the special case of \( f \) linear and \( Y_n - f(X_n) \) normally distributed.

It turns out, however, that the case of normally distributed deviations is the most difficult one. Abdelhamid (1971) and Anbar (1971) studied the effect of using \( X_{n+1} = X_n - an^{-1}h(Y_n) \) if \( Y_n - f(X_n) \) are distributed, conditionally given past, according to a density \( g \) with \( 0 < l(g) = \int (g'/g)g < +\infty \). They found that the optimal \( h \) is, under mild conditions, equal to \( -g'/g \). The result is then that \( n^{-1}(X_n - \theta) \) is asymptotically normal \( (0, \int (f'(\theta))(f'(\theta))^{-2}) \). The remarkable fact is that, with \( h \) optimal, the variance of the asymptotic distribution of \( X_n \) is not only minimal within the class of stochastic approximation methods but is also minimal within the class of all regular unbiased estimators of \( \theta \). The last is true in the sense that the Cramér–Rao bound for the special case of \( f(x) = \alpha(x - \theta) \) is equal to \( \int (f'(\theta))(f'(\theta))^{-2} \). The asymptotic efficiency obtains despite the very simple recurrence relation generating the sequence \( \{X_n\} \).

Received October 6, 1971.

1 Research partially supported by NSF Grant No. GP-23480.

486
The purpose of this paper is to show that the above result can be obtained without the knowledge of \( g \).

The assumptions under which this result is obtained concern the function \( f \) and the distributions of the estimates \( Y_n \) of the function values. Concerning \( f \) (cf. Assumption (2.1)) we assume less than in previous work (cf. e.g. Venter (1967) and Fabian (1968)); see (5.1). Concerning \( Y_n \) we assume, as Abdelhamid (1971) and Anbar (1971) do, that conditional distributions of \( Y_n - f(X_n) \), given the past, are determined by a symmetric density \( g \). We also assume that \( g \) is non-increasing on \([0, +\infty)\), has a continuous derivative \( g' \) and satisfies \( I(g) < +\infty \).

As we mentioned above in a special case \( f(x) \) may be \( \alpha(x - \theta) \) with \( \alpha > 0 \), and if \( \alpha \) is known we may as well assume that \( \alpha = 1 \). Then we may subtract \( X_i \) from the \( Y_i \)'s and obtain observations of \( V_i - \theta, V_2 - \theta, \ldots \) with \( V_i \) independent and distributed according to \( g \). (Conversely we may construct \( Y_i \) from \( V_i - \theta \).) Asymptotically efficient estimators of the location parameter \( \theta \), not requiring the knowledge of \( g \), were given by van Eeden (1970) and Weiss and Wolfowitz (1970a) (the second paper treats also scale parameters; cf. also Weiss and Wolfowitz (1970b)) and it is worthwhile to compare assumptions concerning \( g \). In the former paper, as compared to our assumptions, \( g' \) is not assumed to be continuous but \(-g'/g\) is assumed to be non-decreasing (this implies and is much stronger than our requirement that \( g \) be non-increasing on \([0, +\infty)\)). The latter paper assumes the existence and uniform continuity of \( g'' \) and boundedness from below by a positive constant of \( g \) on an open interval \( I \) such that the measure of the closure of \( I \) under \( g \) is one; the results are formulated for the case where this measure is at least \( 1 - \delta \) when "approximate" asymptotic efficiency is obtained.

The organization of this paper is as follows: Section 2 lists some assumptions, Section 3 contains a preliminary result, Section 4 the main theorem and Section 5 contains remarks and comments.

The author was privileged to have stimulating and fruitful discussions and correspondence with Professor Jack Wolfowitz about the problem at an early stage of the work. The author also benefited from having had access to the results of Anbar (1971) and Abdelhamid (1971) before they were generally available.

2. Basic assumptions and notation. All random variables we shall talk about are supposed to be defined on a probability space \((\Omega, \mathcal{F}, P)\). Relations between random variables, including convergence, are meant with probability one, unless specified otherwise. The real line is denoted by \( R \) and \( \mathcal{B} \) denotes the class of all Borel subsets of \( R \).

We shall list some assumptions for later reference. Only Assumptions (2.1) and (2.2) appear in the final result, Theorem 4.1. Assumptions (2.3), (2.4) and (2.5) are auxiliary.
(2.1) **Assumption.** \( f \) is a function defined on \( R, \theta \in R \), and for every \( \varepsilon > 0 \)
\[
\sup \{ f(x); -\varepsilon^{-1} < x - \theta < -\varepsilon \} < 0, \quad \inf \{ f(x); \varepsilon < x - \theta < \varepsilon^{-1} \} > 0,
\]
f is bounded on bounded intervals and has a derivative in a neighborhood of \( \theta \), which is continuous at \( \theta \) and
\[
f'(\theta) = d > 0.
\]

(2.2) **Assumption.** Assumption (2.1) holds and \( X_1, X_2, \ldots, Y_1, Y_2, \ldots \) are random variables, \( \{ \mathcal{F}_n \} \) a non-decreasing sequence of \( \sigma \)-algebras, each containing the \( \sigma \)-algebra \( \mathcal{F} \{ X_1, \ldots, X_n, Y_1, \ldots, Y_{n-1} \} \) generated by the indicated family of random variables. We suppose that \( Y_n - f(X_n) \) is, conditionally given \( \mathcal{F}_n \), distributed according to a distribution function \( G \) which is symmetric, has zero expectation, has a density \( g \) and a continuous derivative \( g' \) of the density everywhere on \( R \). The density is non-increasing on \([0, +\infty)\) and \( 0 < I(g) = \int (g'/g)^2 \, dG < +\infty \).

(2.3) **Assumption.** Assumption (2.2) holds and \( h_n \) are measurable functions on \( (\Omega \times R, \mathcal{F}_n \times \mathcal{B}) \) such that \( h_n(\omega, \cdot) \) are odd, nonnegative on \([0, +\infty)\) for every \( \omega \). Further \( D_n \) are nonnegative \( \mathcal{F}_n \)-measurable random variables and
\[
|h_n(\omega, t)| \leq n^{1+\varepsilon_1}X_{(-\infty, n)}(t), \quad D_n \leq n^{1+\varepsilon_1}
\]
with a positive \( \varepsilon_1 < \frac{1}{2} \).

We shall write \( h_n(t) \) for \( h_n(\cdot, t) \) and \( h_n(Y_n) \) for \( h_n(\cdot, Y_n(\cdot)) \).

The random variables \( X_1, X_2, \ldots \) satisfy
\[
X_{n+1} = X_n - n^{-1}D_n h_n(Y_n) - n^{-1}(\log n)^{-1+\varepsilon_1}G_n
\]
where
\[
G_n = (Y_n \vee (-y_n)) \wedge y_n
\]
with \( y_n = n^{\varepsilon_1} \) if \( G \) has a finite second moment, \( y_n = (\log n)^{1-2\varepsilon_1} \) otherwise.

(2.4) **Assumption.** Assumption (2.3) holds. For almost all \( \omega, h_n(\omega, \cdot) \to -g'/g \) on the set \( \{ t; g(t) > 0 \} \) and \( D_n \to (I(g)D)^{-1} \).

(2.5) **Assumption.** Assumption (2.4) holds and
\[
\int [h_n(t + \eta_n(t)) + g'/g)^2 \, dG \to 0
\]
for any sequence \( \{ \eta_n \} \) of functions on \( \Omega \times R \) such that, for almost all \( \omega, h_n(\omega, t + \eta_n(\omega, t)) \) are Borel measurable with respect to \( t \) and \( |\eta_n| \leq |f(X_n)| \).

3. Preliminary results on convergence of \( X_n \).

(3.1) **Theorem.** If Assumption (2.3) holds then \( (\log n)^{\delta}(X_n - \theta) \to 0 \) for every \( \beta > 0 \). If Assumption (2.4) holds then \( n^{\beta}(X_n - \theta) \to 0 \) for every \( 0 < \beta < \frac{1}{2} - 2\varepsilon_1 \).

If Assumption (2.5) holds then \( n^\beta(X_n - \theta) \) is asymptotically normal with zero mean and variance \( d^{-1}I^{-1}(g) \).

**Proof.** Without loss of generality we may assume that \( \theta = 0 \).
(i) Suppose Assumption (2.3) holds. Notice that $E_{\omega_n} h_n(Y_n) = \Psi_n(f(X_n))$ where

$$\Psi_n(\Delta) = \int h_n(t + \Delta)g(t)\,dt = \int_0^{\infty} h_n(t)[g(t - \Delta) - g(t + \Delta)]\,dt$$

the last representation following from the fact that $h_n$ is odd and $g$ symmetric. Since $g(t - \Delta) - g(t + \Delta) = g(\Delta - t) - g(\Delta + t)$ and since $g$ is non-increasing on $[0, +\infty)$, $h_n$ nonnegative, we conclude that the integrand in the second integral in (1) is nonnegative for $\Delta > 0$. Since $\Psi_n$ is odd, we obtain, for all $\Delta$,

$$\Delta \Psi_n(\Delta) \geq 0.$$ 

Next $\tilde{Y}_n = q_n(Y_n)$ with $q_n(y) = (y \vee (-y_n)) \wedge y_n$ and $E_{\omega_n} \tilde{Y}_n = \varphi_n(f(X_n))$, $\varphi_n(\Delta) = \int q_n(t + \Delta)g(t)\,dt$. Differentiating,

$$\varphi_n'(\Delta) = \int_{\Delta}^{\infty} q_n(t)g(t)\,dt.$$ 

Writing now (2.3.2) as $X_{n+1} = X_n - U_n$ we obtain

$$X_n E_{\omega_n} U_n \geq n^{-1}(\log n)^{-1+\epsilon} X_n \xi_n,$$

with $\xi_n = E_{\omega_n} \tilde{Y}_n = f(X_n)\varphi_n'(\Delta_n)$, \(|\Delta_n| \leq |f(X_n)|\). Also, easily,

$$E_{\omega_n} U_n \leq C n^{-1-4\epsilon},$$

with a constant $C$.

Set $h(x) = \frac{1}{2}x^2$, $N_n = E_{\omega_n} U_n$, $B_n = (X_n \xi_n)^i$ (notice that $X_n \xi_n \geq 0$ according to (2.1.1) and (3)), $\alpha_n = n^{-1}(\log n)^{-1+\epsilon}$. Then (4) gives

$$h'(X_n)N_n \geq \alpha_n B_n^2.$$ 

Setting $\beta_n = C n^{-1-4\epsilon}$, $\gamma_n = \xi_n = 0$ we have $\sum \alpha_n = +\infty$, $\sum \beta_n < +\infty$ and an application of Lemma (3.3) in Fabian (1971) or Theorem 5.2 in Fabian (1960) implies that \(|h(X_n)|\) converges to a finite random variable and $B_{n+1} \rightarrow 0$ for a subsequence \(\{B_n\}\). If $\omega$ is a point in $\Omega$ at which the two properties hold then $\Delta_n(\omega)$ is a bounded sequence as $f$ is bounded on bounded intervals, $\varphi_n'(\Delta_n(\omega)) \rightarrow 1$, and (2.1.1) implies $X_n(\omega) \rightarrow 0$. This in turn implies that the limit of $h(X_n(\omega))$, which exists, must be 0. Thus $X_n \rightarrow 0$.

By Assumption (2.1), $f'(0)$ exists and thus

$$f(X_n) = d_n X_n, \quad d_n \rightarrow d$$

where $d_n$ are $\mathcal{F}_n$-measurable random variables. Then

$$N_n = n^{-1}D_n d_n X_n d_n X_n^{-1} \Psi_n(d_n X_n) + n^{-1}(\log n)^{-1+\epsilon}\varphi'(\Delta_n)d_n X_n.$$

We have already established $\Delta^{-1} \Psi_n(\Delta) \geq 0$ for every $\Delta$, and $\varphi_n'(\Delta_n) \rightarrow 1$.

It is easy to verify

$$\Delta^{-1} \Psi_n(\Delta) \leq C_1 n^\epsilon$$

for all $\Delta \neq 0$ and a constant $C_1$ by differentiating $\Psi_n(\Delta) = \int h_n(t)g(t - \Delta)\,dt$ and using the fact that $l(g) < +\infty$ implies the integrability of $g'$, and (2.3.1). Thus,
if \( \varepsilon_0 < \varepsilon_1 \),

\[
N_n = n^{-\gamma_0} \gamma_0 n^{-\gamma_1} x_n
\]

with \( \gamma_0 \to +\infty \), \( \gamma_\infty \leq C_2 n^{3\varepsilon_1} \) and a constant \( C_3 \). Eventually, depending on \( \omega \),

\[
0 < 1 - n^{-1}(\log n)^{-1+\gamma_0} n^{1+\varepsilon_0} \leq 1 - n^{-1}(\log n)^{-1+\varepsilon_0} \quad \text{and} \quad
\]

\[
X_{n+1}^2 \leq (1 - n^{-1}(\log n)^{-1+\varepsilon_0}) X_n^2 - 2 V_n + W_n
\]

with

\[
V_n = (X_n - N_n)(U_n - N_n), \quad W_n = (U_n - N_n)^2.
\]

Thus

\[
V_n = (1 - n^{-1}(\log n)^{-1+\varepsilon_0}) X_n(U_n - N_n).
\]

Next we want to show that if \( \beta_n \) are positive numbers, \( \beta_n \leq n^{2\beta} \) with \( 0 < \beta < \frac{1}{2} - 2\varepsilon_1 \), then

\[
\sum_{n=1}^{\infty} \beta_n W_n < \infty, \quad \sum_{n=1}^{\infty} \beta_n V_n < \infty
\]

on the set \( \{ \beta_n n^{-2\beta} X_n^2 \to 0 \} \).

The convergence of the first series follows from the fact that it has a finite expectation as

\[
E_{\mathcal{F}_n} W_n \leq C_3 n^{-2+4\varepsilon_1}
\]

with a constant \( C_3 \). Concerning the second series, with \( \eta = 1 - 4\varepsilon_1 - 2\beta \),

\[
E_{\mathcal{F}_n} \beta_n^2 V_n^2 \leq C_3 \beta_n^2 X_n^2 n^{-2+4\varepsilon_1} \leq C_3 \beta_n^2 n^{-2\beta} X_n^2 n^{-1-\eta}
\]

with the last term summable on the set indicated in (13). The convergence of

\[
\sum \beta_n V_n
\]

on this set then follows by the generalized Borel–Cantelli lemma (Lemma 10, Dubins and Freedman (1965)).

Now set \( \beta_n = (\log n)^\alpha \) with a \( b > 0 \), \( \alpha_n = \beta_{n+1}/\beta_n \), verify that

\[
\alpha_n \leq 1 + b(n \log n)^{-1}
\]

\[
\alpha_n(1 - n^{-1}(\log n)^{-1+\varepsilon_0}) \leq 1,
\]

eventually

and, from (11),

\[
\beta_{n+1} X_{n+1}^2 \leq \beta_n X_n^2 - 2 \beta_{n+1} V_n + \beta_{n+1} W_n.
\]

By (13) (take any positive \( \beta < \frac{1}{2} - 2\varepsilon_1 \) to obtain \( \beta_{n+1} n^{-2\beta} X_n^2 \to 0 \)) the terms \( \beta_{n+1} V_n, \beta_{n+1} W_n \) have convergent sums and thus \( \beta_n X_n^2 \) is bounded. Since \( b > 0 \) was arbitrary the proof of the first part of the theorem is completed.

(ii) Suppose Assumption (2.4) holds. Express the nonnegative (for \( t \geq 0 \)) difference \( g(t - \Delta) - g(t + \Delta) \) in (1) as \( 2\Delta(-g'(\eta_n)) \) with \( |\eta_n - t| < \Delta \). If we let \( \Delta = d_n X_n, \eta_n \) depends on \( \omega \) and \( t \) and

\[
\mathcal{W}(d_n X_n) = 2d_n X_n \int_0^\infty h_n(t)(-g'(\eta_n)) \ dt.
\]

As we noticed the integrands are nonnegative, converge (for almost all \( \omega \)) pointwise on \( \{ t; g(t) > 0 \} \) to \( (-g'/g)(-g') \) by Assumption (2.4) and since \( g' \) is con-
tinuous. Using the Fatou lemma gives
\[
\liminf \int_0^\infty h_n(t)(-g'(\gamma_n)) \, dt \geq \int_{\{t; t \geq 0, g'(t) > 0\}} \left( \frac{g'}{g} \right)^2 \, dG = \frac{1}{2} I(g)
\]
and thus, if we interpret $\Delta^{-1} \Psi_n(\Delta) = I(g)$ for $\Delta = 0$,
\begin{equation}
(16) \quad \liminf (d_n X_n)^{-1} \Psi_n(d_n X_n) \geq I(g).
\end{equation}
Thus from (8) we obtain a strengthening of (10), namely
\begin{equation}
(17) \quad N_n = n^{-1} \kappa_n X_n
\end{equation}
with
\begin{equation}
(18) \quad \liminf \kappa_n \geq 1, \quad \kappa_n \leq C_t n^{2z}\n\end{equation}
with a constant $C_t$. From here (11) can be strengthened to
\begin{equation}
(19) \quad X_{n+1}^2 \leq (1 - 2n^{-1} \kappa' n) X_n^2 - 2V_n + W_n
\end{equation}
where $\kappa' = \kappa_n \to 0$.

Suppose $n^{-\beta_0} X_n^2 \to 0$ for a $0 \leq \beta_0 < \frac{1}{2} - 2\varepsilon_1$, which we know is true at least if $\beta_0 = 0$. Take a $\beta < \frac{1}{2} - 2\varepsilon_1, \beta > \beta_0$ and write $\beta^2 n^{-1} \Phi_n^2$ as $\beta^2 n^{-2} \Phi_n^2 n^{2\beta_0} X_n^2$ to see that this sequence converges to zero if $\beta_n = n^{-\beta + \beta_0}$. Then, by (13),
\[ \sum_{n=1}^\infty \beta_n W_n < +\infty \text{ and } \sum_{n=1}^\infty \beta_n V_n < +\infty. \]
Repeating an argument from part (i), or directly using Lemma 4.3 in Fabian (1967) yields the boundedness of $n^{1+\beta_0} X_n^2$.

By induction, $n^{1+\beta} X_n^2 \to 0$ for every $\beta < \frac{1}{2} - 2\varepsilon_1$.

(iii) Suppose Assumption (2.5) holds. As in (ii),
\[
(d_n X_n)^{-1} \Psi_n(d_n X_n) = -\int_{-\infty}^{+\infty} h_n(t)g'(t - \theta_n) \, dt = -\int_{-\infty}^{+\infty} h_n(t + \theta_n) \frac{g'}{g}(t) \, dG(t)
\]
with $|\theta_n| < |d_n X_n|$. Using the Schwarz inequality and (2.5.1) we obtain
\[
[(d_n X_n)^{-1} \Psi_n(d_n X_n)]^2 \leq \int h_n^2(t + \theta_n) \, dG(t) \int \left( \frac{g'}{g} \right)^2 \, dG \to P(g);
\]
this with (16) gives
\begin{equation}
(20) \quad (d_n X_n)^{-1} \Psi_n(d_n X_n) \to I(g)
\end{equation}
and from (8)
\begin{equation}
(21) \quad N_n = n^{-1} \gamma_n X_n, \quad \gamma_n \to 1.
\end{equation}
Next, denoting the conditional variance, given $\mathcal{F}_n$, by $\text{Var}_{\mathcal{F}_n}$, we have
\[
\text{Var}_{\mathcal{F}_n} h_n(Y_n) = \int h_n^2(t + d_n X_n) \, dG(t) - \Psi_n^2(d_n X_n) \to I(g)
\]
since $h_n(t + d_n X_n)$ converge to $-g'/g$ in $L_d(g)$ and $\Psi_n(d_n X_n) \to 0$. Then
\begin{equation}
(22) \quad D_n^2 \text{Var}_{\mathcal{F}_n} h_n(Y_n) \to d^{-2} I(g)^{-1}.
\end{equation}
Consider now $(\log n)^{-1+\gamma_n} Y_n$ and its conditional, given $\mathcal{F}_n$, variance. If $y_n$ in (2.3.3) are $(\log n)^{1-2z}$, this conditional variance is bounded by $(\log n)^{-2z}$. If $y_n = n^{1+}$ then $G$ has a finite variance, say $\sigma^2$, and the conditional variance of $Y_n$
will be less or equal to \( \sigma^2 \) on the set where \(|E_{\mathcal{F}_n} \tilde{Y}_n| = |d_n X_n \varphi_n'(\Delta_n)| \leq n^t\) where \( \varphi_n'(\Delta_n) \to 1 \). Since this will eventually happen, we obtain, under both choices of the \( \gamma_n \)'s that

\[
\text{Var}_{\mathcal{F}_n} (\log n)^{-1+\epsilon} \tilde{Y}_n \to 0.
\]

The two random variables, \( h_n(Y_n) \) and \((\log n)^{-1+\epsilon} \tilde{Y}_n\) are not independent, but by Minkowski inequality it follows from (22) and (23) that

\[
n^2 E_{\mathcal{F}_n}(U_n - N_n)^2 \to d^{-2} I(g)^{-1}.
\]

Suppose now \( r > 0 \), forget the old meaning of \( V_n \) and set \( V_n = n(U_n - N_n) \). Notice that \(|V_n| \leq n^2 C_0\), with a constant \( C_0 \), so that \(|V_n^2 \geq r n|\) is eventually an empty set and

\[
E V_n^2 \chi_{|V_n^2 | \geq r n} \to 0.
\]

Summarizing, we have

\[
X_{n+1} = (1 - n^{-1} \gamma_n) X_n - n^{-1} V_n;
\]

the measurability properties of \( \gamma_n \), (21), (24) and (25) imply, by Theorem 2.2 in Fabian (1968), the last part of our theorem.

4. The main result.

(4.1) **Theorem.** Suppose Assumptions (2.1) and (2.2) hold with \( \mathcal{F}_n \) as defined below. (The requirement in Assumption (2.2), that \( \mathcal{F}_n \supset \mathcal{F}\{X_1, X_2, \ldots, Y_1, \ldots, Y_{n-1}\} \), will be automatically satisfied.) Suppose \( \{m_i\} \) is an increasing sequence of positive integers such that \( |m_i| \to 0 \) and \( \{U_i\} \) is a sequence of random variables such that, with

\[
\mathcal{F}_n = \mathcal{F}\{X_1, Y_1, \ldots, X_{n-1}\} \cup \{U_i; m_i < n\},
\]

\[
E_{\mathcal{F}_{m_i}} U_i = (2c_i)^{-1}[f(X_{m_i} + c_i) - f(X_{m_i} - c_i)],
\]

\[
E_{\mathcal{F}_{m_i}} (U_i - E_{\mathcal{F}_{m_i}} U_i)^2 \leq c_i^{-2} C
\]

with a constant \( C \) and \( c_i \) of the form

\[
c_i = c I^{-\gamma}, \quad c > 0, \quad 0 < \gamma < \frac{1}{2}.
\]

Then the sequence \( X_n \), as defined by the procedure described below, converges to \( \theta \) and \( t_n^{-1}(X_n - \theta) \) is asymptotically normal \( (0, d^{-2} I^{-1}(g)) \), where \( t_n = n + 2 \text{ card } \{l; m_i < n\} \) is the number of observations used to construct \( X_n \).

(4.2) **The Procedure.**

(a) **Estimation of \( d \).** Set \( \hat{U}_n \) equal to the arithmetic mean of all \( U_i \) with \( m_i < n \) and set

\[
u_n = (0 \lor \hat{U}_n).
\]

(b) **The sequence \( \{D_n\} \).** Denote by \( G_n \) the empirical distribution function of \( Y_1, Y_2, \ldots, Y_n \), set \( \varepsilon_n \geq (\log n)^{-\beta_0} \) for a \( \beta_0 > 0 \), \( \varepsilon_n \to 0 \), and select positive \( \Delta_n, \delta_n \)

\[
\varepsilon_n \geq (\log n)^{-\beta_0} \text{ for } \beta_0 > 0, \varepsilon_n \to 0, \text{ and select positive } \Delta_n, \delta_n
\]
such that

\[(2) \quad \Delta_n \to 0, \quad \varepsilon_n \Delta_n^{-1} \to 0, \quad \delta_n \varepsilon_n^{-1} \to 0, \quad n^{-r} \delta_n^{-1} \varepsilon_n^{-1} \to 0 \quad \text{for an } r < \frac{1}{2}.
\]

Use the symbol \(q'(x)\) for \(q(x + c) - q(x - c)\) and set

\[(3) \quad h_n(t) = \frac{(G_n, s)^{\delta_n(t)}}{2 \delta_n G_n^{1/2}(t)} \chi_{\{\sigma_n \to \infty\}}(G_n, s(t))
\]

for all \(t\) in \(T = \{(2j - 1)\Delta_n; j = 0, 1, -1, \ldots\}\) and let \(h_{n+1}\) be constant on the intervals \((2j - 2)\Delta_n, 2j\Delta_n\). Set

\[(4) \quad D_n = \left\{ u_n \frac{\partial}{\partial u_n} \right\}^{(\gamma_n)} dG_n^{\frac{1}{2}} \cap n^{-1}.
\]

(c) Choice of functions \(h_n\). Choose \(h_n^0\) to satisfy conditions (b) but with \(\varepsilon_n \geq n^{-\beta_1}\) with \(0 < \beta_1 < \frac{1}{2} - 2\varepsilon_1\) (it may be the same, or different, choice of \(h_n^0\) than in (b)). Set

\[h_n(t) = \left( \begin{array}{c} \frac{1}{2} h_n^0(t) - h_n^0(-t) \end{array} \right) \cap \left( n^{\gamma_1} \chi_{\{\sigma_n \to \infty\}}(t) \right) \quad \text{for } t \geq 0
\]

\[= -h_n(-t) \quad \text{for } t < 0.
\]

(d) The recurrence relation for \(X_n\) is (2.3.2).}

(4.3) PROOF OF THEOREM (4.1). We shall prove the theorem by verifying Assumptions (2.1) to (2.5).

(i) Assumptions (2.1), (2.2) are required in our theorem. The measurability conditions on \(D_n\) and \(h_n\) as well as (2.3.1) are obvious from the definitions of \(D_n\), \(h_n\) and (2.3.2) holds by assumption. Thus Assumption (2.3) holds. Theorem (3.1) implies \((\log n)^{\gamma} (X_n - \theta) \to 0\) for every \(\beta > 0\).

(ii) Refer by 1 to Fabian (1973). We shall use Theorem (I.2.2). Conditions imposed there and in Extension (I.2.4) on \(G\) are repeated in Assumption (2.2).

If we put \(Z_i = f(X_i), V_i = Y_i - Z_i\) then \(V_n\) is distributed, conditionally given \(Z_1, \ldots, Z_n, V_1, \ldots, V_{n-1}\), according to \(G\). Condition (I.2.2.1) is repeated in (4.2.2). We then obtain from Extension (I.2.4) that \(h_n(\omega, \cdot) \to g'g\) for almost all \(\omega\) on \([t; g(t) > 0]\); the truncation of our \(h_n^0\) by \(n^{\gamma_1} \chi_{\{\sigma_n \to \infty\}}\) obviously does not affect this property. For \(h_n^0\) as defined in (4.2.b) we have \(\varepsilon_n n \uparrow +\infty\), \(\sum_i n^{-i} \varepsilon_i^{-1} |Z_n| < +\infty\) as \((\log n)^{\gamma+i} |Z_n| \to 0\) and the Kronecker lemma implies (I.2.2.4). It follows then by Theorem (I.2.2) that \(\frac{\partial}{\partial u_n} dG_n^{\frac{1}{2}} \to l(g)\).

To show that Assumption (2.4) holds it remains only to prove that \(u_n\), or for that matter, \(\bar{U}_n\), converge to \(d\). Set \(W_i = U_i - E_{\mathbb{E}^m} U_i\). Then \(W_i\) is an orthogonal sequence,

\[\sum_{i=1}^n (\log l)^{h-2} E W_i^2 \leq c^{-2} \sum_{i=1}^n (\log l)^{h-2} \to +\infty
\]

and Theorem 4.5.2 in Doob (1953) implies that \(l^{-1} \sum_i W_i \to 0\). Since \(E_{\mathbb{E}^m} U_i = f(X_{m_i} + \eta_i)\) (with \(\eta_i \to 0\)), eventually, depending on \(\omega\), we obtain, using Assumption (2.1), \(E_{\mathbb{E}^m} U_i \to f'(\theta)\) and \(\bar{U}_n \to d\). This means that Assumption (2.4) holds and Theorem (3.1) implies \(n^{\gamma} (X_n - \theta) \to 0\) for every \(\beta < n^{\gamma+i}\).

(iii) The \(\varepsilon_n\)'s from (4.2.c) again satisfy all the requirements of Theorem (I.2.2).
Condition (1.2.2.4) holds since $n\delta|Z_n| \to 0$ with a $\beta > \beta_1$ and $(n\delta)^{-1} \sum_{n=1}^{\infty} |Z_j| \leq n^{-1+\delta_1} \sum_{n=1}^{\infty} n^{-\delta} n^\delta |Z_j| = O(n^{\delta_1-\delta})$. Also (1.2.3.1) is satisfied as $\gamma_n \leq |Z_n|$ in Assumption 2.5 and $\varepsilon_n^{-1} \gamma_n \to 0$. By Extension (1.2.4) the assertion in Extension (1.2.3) then implies (2.5.1). This means that Assumption (2.5) holds and by Theorem (3.1) $X_n$ has properties claimed in our theorem since $\tau_n/n \to 1$ because $l/m \to 0$.

5. Remarks and comments.

(5.1) On Assumption (2.1). In all previous work $f$ is assumed to satisfy $|f(x)| \leq A + B|x|$ for some constants $A, B$. Here the truncation of the $Y_n$'s makes this condition unnecessary. The main reason to use truncation was, however, to avoid a similar requirement for the conditional expectation of $-\langle g'/g \rangle(Y_n)$. If $f$ satisfies the above condition and $G$ has a finite second moment, $Y_n$ can be used in (2.3.2) instead of $\hat{Y}_n$; the required modification of proofs is slight.

Conditions under which the optimum constant $\alpha = d^{-1}$ for the coefficients $an^{-1}$ was previously estimated, included the requirement of a bounded second derivative in a neighborhood of $\theta$ (Fabian (1968); more stringent conditions in Venter (1967)). The weakening of this condition here was made possible by another, simpler, method of estimation of $d$. Actually the condition on $f'$ can be reduced still further to the only requirement of $f'$ existing at $\theta$. Indeed this is enough for (3.1.7) which is the first instant of the use of $f'$. The second time the properties of $f'$ are used is in (4.3.ii) to prove that $E_{C_{m1}}(U_j) = (2c_i)^{-1}[f(X_{m1} + ci) - f(X_{m1} - ci)] \to d$. This could be still obtained under the mere assumption of the existence of $f'$ at $\theta$ if the $c_i$ are chosen to be converging to 0 but $c_i \geq (\log m_i)^{-\beta}$ for some $\beta > 0$. In this case $f(h) = dh + o(h)$, thus $E_{C_{m1}}(U_j) = (2c_i)^{-1}(X_{m1} + ci) + o(c_i^{-1}(X_{m1} + ci)) = d + o(1)$.

(5.2) On Assumption (2.2). The usual assumption was that $E_{C_{m1}} Y_n^2 \leq \sigma^2$ at least for $X_n$ near to $\theta$. The truncation makes it possible to dispose of this assumption, but then the truncation of the $Y_n$'s in (2.3.2) has to be more severe. However the main components $h_n(Y_n)$ in the right-hand side of (2.3.2) have bounded variances, at least when $X_n$ is near to $\theta$ (cf. (3.1.22)).

(5.3) On estimation of $-g'/g$ and $I(g)$. Of course if we could we would use the formula

$$X_{n+1} = X_n + \frac{1}{n d(g)} (g'/g)(Y_n).$$

Estimation of $d$ is easy. To establish a convergence of type $n^\delta(X_n - \theta)$ we may overestimate $-g'/g$ in the sense of (3.1.16) but we must not underestimate $I^{-1}(g)$. That explains why there is a wider choice of constants in estimating $-g'/g$ than in estimating $I(g)$ as given by the condition $\varepsilon_n \leq (\log n)^{-\beta}$ in (4.2.3) and condition $\varepsilon_n \geq n^{\delta_1}$ with $\beta_1 < \frac{1}{3} - 2\epsilon$, in (4.2.4). The proof could be somewhat simplified if we did not want to show the possibility of this wider choice of $\varepsilon_n$ for estimating $-g'/g$. 
(5.4) On the choice of truncation. Obviously truncation at $n^t$, at various places, or by $\chi_{(-\infty,n]}$ as in (2.3.1) was chosen quite arbitrarily and the function $n^t$ can be replaced by any other function which increases sufficiently slowly. The function $\chi_{(-\infty,n]}$ can be replaced by $\chi_{(-v(n),v(n)]}$ with any $v(n) \to +\infty$.

(5.5) Computational aspects. Introducing the $h_n$ into the recurrence formula for $X_n$ destroys the extreme simplicity of the original stochastic approximation procedure. However, it is obvious that all the convergence properties of $h_n$ are shared by any subsequence $h_{n_t}$ and then by the sequence $\bar{h}_n = h_{n_t}$ for $n_t \leq n < n_{t+1}$. This makes it possible to compute a new estimate of $-g'/g$ only once in a while.

REFERENCES


Department of Statistics and Probability
Michigan State University
Wells Hall
East Lansing, Michigan 48823